Best Practice for Caching of Single-Path Code

Schoeberl, Martin; Cilku, Bekim; Prokesch, Daniel; Puschner, Peter

Published in:
Proceedings of 17th International Workshop on Worst-Case Execution Time Analysis

Link to article, DOI:
10.4230/OASIcs.CVIT.2016.23

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Best Practice for Caching of Single-Path Code*

Martin Schoeberl[1], Bekim Cilku[2], Daniel Prokesch[2], and Peter Puschner[2]

1 **Department of Applied Mathematics and Computer Science**
Technical University of Denmark
masca@imm.dtu.dk
2 **Institute of Computer Engineering**
Vienna University of Technology, Austria
{bekim,daniel,peter}@vmars.tuwien.ac.at

Abstract

Single-path code has some unique properties that make it interesting to explore different caching and prefetching alternatives for the stream of instructions. In this paper, we explore different cache organizations and how they perform with single-path code.

1998 ACM Subject Classification C.3 Real-Time and Embedded Systems

Keywords and phrases single-path, method cache, prefetching

Digital Object Identifier 10.4230/OASIcs.CVIT.2016.23

1 Introduction

Worst-case execution time (WCET) analysis is a non-trivial analysis problem. It becomes especially difficult with more complex processor architectures. A strategy to simplify WCET analysis is to write programs that have a constant execution time, i.e., the best-case and worst-case execution time are equal. In that case, we do not need to analyze the program, but can simply measure the execution time. Single-path code gives constant execution time.

Single-path code is code that is structured so that there are no data dependent control flows. On an `if/else` condition both conditions are executed. However, to retain the program's semantics and data flow, all instructions are executed with a predicate. The compiler sets these predicates according to the original conditions of the branching code. When executing single-path code, instructions whose predicate evaluates to `false` do not update the processor state, i.e., they act as `nop` instructions. Loops always execute the maximum number of iterations (their so-called loop bound), which is a known number in a real-time context. Like the `if/else` case, the original loop condition is used to evaluate to a predicate and all instructions within loops are predicated.

Single-path code can be manually coded or a compiler can translate *normal* code to single-path code. The translation of an `if/else` condition is also a common technique in VLIW compiler applied for small code fragments to avoid expensive branches. This is called if conversion [1].

The time-predictable execution of single-path code demands two features from a processor: (1) the processor needs to support predicates or a conditional move and (2) a predicated

* This paper was partially funded by the EU COST Action IC1202: Timing Analysis on Code Level (TACLe) and the European Union's 7th Framework Programme under grant agreement no. 288008: Time-predictable Multi-Core Architecture for Embedded Systems (T-CREST).

instruction shall have the same execution time irrespective of whether the predicate evaluates to `true` or `false`. Patmos fulfills both conditions.

Patmos contains also a special instruction cache that caches full functions. For historical reasons this cache is named method cache (it appeared first in a Java processor). Cache misses can only occur at function calls or returns. Caching full functions has one drawback: code that is not executed is still loaded into the cache. However, as programs organized as single-path code execute all their instructions, this main drawback disappears. Therefore, our hypothesis is that the method cache is a good cache organization for single-path code.

This paper explores the method cache in the context of single-path code. We compare and evaluate the method cache against a standard instruction cache using the TACLeBench benchmarks [6]. Furthermore, we explore performance benefits of an extension of a standard instruction cache with a prefetcher that has been especially designed for single-path code.

The paper is organized in 6 sections: The following section presents related work. Section 3 provides background on single-path code generation and the time-predictable Patmos processor. Section 4 describes different options of caching for single-path code. Section 5 evaluates the different caching options on the Patmos processor and compares them. Section 6 concludes the paper.

2 Related Work

For real-time systems, caches are also one of the main sources of temporal uncertainty. State-of-the-art cache analysis tools are using abstract interpretation for classifying cache accesses and with that also the predictability of the cache behavior [12]. However, even if these approaches derive safe bounds, the precision of the results derived from the abstracted models strongly vary depending on the cache architecture and replacement policy [9]. For example, an abstract model for the LRU replacement policy achieves better predictability than a model for FIFO or PLRU [17].

Another mechanism that aims at making caches more predictable is cache locking [14]. This technique loads memory contents into the cache and locks it to ensure that it will remain unchanged afterwards. The benefit of cache locking is that all accesses to the locked cache lines will always result into cache hits. The cache content can be locked entirely [7] or partially, it can be locked for the whole system lifetime (static cache locking) or it can be changed at runtime (dynamic cache locking) [5]. Although cache locking increases predictability, it reduces performance by restricting the temporal locality of the cache to a set of locked cache lines.

In contrast to conventional code, single-path conversion overcomes predictability issues by generating code that has only a single trace of execution. Thus, keeping traces of possible cache states is no more needed. Furthermore, the use of single-path code eliminates the necessity for cache locking.

3 Background

This paper builds on prior research work on single-path code and research on the time-predictable computer architecture developed for the T-CREST platform.

3.1 Single-path Code Generation

Puschner and Burns propose single-path code to simplify WCET analysis by avoiding data-dependent control flow decisions [15]. The defining property of single-path code is that any

execution follows a single instruction trace, independent from input data. This is achieved by conversion of control dependence to data dependence, with the use of predicated instructions. In code that is WCET analyzable, loops must be bounded. The compiler transforms input-data dependent loops such that they iterate for a fixed number of times, which is the local loop bound [13].

Single-path code generation provides a constructive approach to predictable real-time code. On a "well-behaved" hardware platform, the execution time for single-path tasks is constant. In this ideal case, WCET analysis simplifies to measurement.

One requirement is that the instruction timing is independent of the instruction operands. Memory accesses introduce another source of variability in execution time. Though, the single-path property makes the code easier to analyze with regards to instruction memory. Abstract interpretation based analysis becomes superfluous, there is no need for approximation. The known singleton instruction stream can be directly applied to a hardware model of the instruction cache (as in simulation). This knowledge is exploited to implement perfectly accurate prefetching schemes for instructions [2].

Data accesses are also subject to execution-time variability. Enforcing local availability of the required data during the task execution may alleviate the problem, e.g., by data cache locking or usage of a scratchpad memory. However, we restrict ourselves to the instruction cache in this paper.

3.2 Patmos and the T-CREST Platform

We explore instruction caching options on the Patmos processor [20], which itself is part of the T-CREST multicore platform [18]. The T-CREST platform aims to build a processor, network-on-chip, and compiler toolchain [16] to simplify WCET analysis. We optimized all components to be time-predictable, even when average-case performance is reduced. AbsInt aiT [8] static WCET analyzer supports the Patmos processor. T-CREST also includes the research WCET analyzer platin [11].

Patmos is a RISC architecture supporting dual-issue instructions. As far as we know, Patmos is timing anomaly free. There is no timing dependency between any two instructions. Even all cache misses (instruction or data) happen in the same pipeline stage (the memory stage). Therefore, only a single cache miss can happen any clock cycle. Patmos uses special forms of instruction and data cache that shall simplify cache analysis. For instructions, Patmos has a method cache [4], which caches whole functions. Besides these special caches Patmos also supports a standard instruction cache, a standard data cache, and instruction and data scratchpad memories.

One issue with a method cache is that full functions are loaded into the method cache, even when only part of it is executed. We attack this issue by splitting larger functions into smaller subfunctions [10]. However, with single-path code there is no code that is not executed. The processor executes all instructions of a called function. Therefore, a method cache may well fit for caching single-path code.

We extended a standard instruction cache by a prefetching unit [3] to improve single-path execution time. This prefetcher only prefetches instructions when the main pipeline will not cause an instruction cache miss.

4 Caching of Single-Path Code

Single-path code is instruction-cache friendly as all instructions that are loaded into the cache are executed, except at the end of a function.

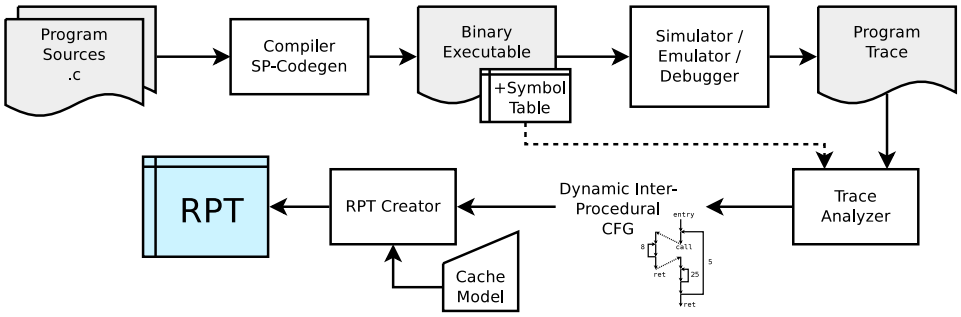

Figure 1 Generation of the Reference Prediction Table (RPT).

4.1 Standard Instruction Cache

A standard instruction cache is organized in cache blocks and can be configured as direct mapped cache or set associative cache. One advantage of using direct-mapped caching for single-path code is the ability of the cache to reduce the miss rate even further when a single-path loop is larger than the cache [2]. For example, if a loop has a size of six cache lines and the cache consists of four cache lines, then after the first iteration the first two lines of the cache will be in conflict and be replaced interchangeably while the third and fourth line will stay unchanged, thus performing as the cache would have a cache lock mechanism. If the same loop is executed on the cache with the same size but is organized as a set associative, then the conflict will appear for every cache line.

4.2 Method Cache

The method cache is an instruction cache designed to simplify WCET analysis. The method cache caches full functions/methods. Therefore, a cache miss can only happen on a call or a return. All other instructions are guaranteed hits and cache analysis can ignore those. Method cache analysis only needs to consider functions and not individual instructions.

One disadvantage of the method cache is that instructions in a function that are not executed are still loaded on a cache miss. However, with single-path code all instructions of a function are always executed. Therefore, the method cache should perform well with single-path code.

4.3 Time-predictable Prefetcher with a Standard Cache

The time-predictable prefetcher exploits properties of single-path code to anticipate future instruction cache accesses to bring those instructions into the cache before they are executed [3]. Correct prediction of the prefetch addresses not only improves execution performance, but also prevents the cache content from pullition of unused instructions.

For higher efficiency, the prefetcher implements an algorithm that prefetch both sequential and non-sequential streams of execution. Anticipating the address of the next sequential prefetching is easy process, since the target address is just the next cache line. Non-sequential prefetching is a harder problem. In such cases, the prefetcher needs to know in advance the outcome of control-flow instructions, to calculate the address of the target that should be prefetched.

A *Reference Prediction Table* (RPT) directs the prefetcher. The entries of the RPT control the behavior of the prefetcher. They contain addresses at which the prefetcher should switch between sequential and non-sequential prefetching. Figure 1 shows the generation of

the RPT. It begins with obtaining the execution trace of the single-path code. We use the Patmos simulator to export the program counter values during a program run. We extract the start addresses of the functions from the symbol table of the executable. The trace analyzer uses the trace and the start addresses to produce a dynamic control-flow graph of the single-path function, where nodes are addresses of single instructions. The trace analyzer identifies call sites, loops, loop nests, and loop iteration counts. The RPT creator then creates entries containing an address that should trigger a change in the behavior of the prefetcher, a destination where to continue prefetching and additional information depending on the entry type. The RPT is a projection of the single-path program which captures its control-flow in units of memory blocks that fit into a cache line.

5 Evaluation

We evaluate the program performance of single-path code with different caching methods. For the comparison, we use the Patmos processor. We configure Patmos for the Altera DE2-115 FPGA board, which means that the main memory is a 16-bit SRAM. This memory results in 21 clock cycles for a burst of 4 32-bit words to fill or spill a 16-byte cache line. All standard caches have the line size of the burst length, 16 bytes. We configure the instruction or method cache to be 8 KB large and the method cache to cache up to 16 functions. The data cache is 4 KB and the stack cache 2 KB. We use hardware simulation to get cycle accurate measurements.

For the evaluation, we use the TACLeBench benchmark collection [6] in version 1.9. We have added an attribute to the benchmark's main function to avoid inlining of this function. Otherwise we did not touch the source of TACLeBench. This main function is also the root function for the single-path code generation. We measure the execution time of the whole program, including initialization and result comparison code, in clock cycles.

We used a subset of the benchmarks. The variation of the execution time of the benchmarks is high, i.e., between hundreds and a billion clock cycles. For practical reasons, we did not use the long running benchmarks, as cycle accurate hardware simulation is time consuming.[1] Furthermore, we dropped benchmarks where we cannot generate single-path code, e.g., recursive benchmarks. Furthermore, we removed two outliners (`ludcmp` and `minver`) as their results showed improvements of factors 3 to 4 for the method cache compared to an instruction cache.

5.1 Baseline

As a baseline, we show the performance difference between using a method cache and a direct mapped instruction cache on normal compiled code. Figure 2 shows the execution time relation between those two configurations (normalized to the execution time with the standard cache). Those measurements are average case measurements and cannot be an indication of WCET analysis bounds. In these average case measurements, we see that some benchmarks perform equally for the two cache configurations. We assume those cases are when the benchmark fits entirely into the cache. Several benchmarks perform better with a normal instruction cache than with the method cache. However, this is an average case measurement and the method cache was designed to simplify WCET analysis.

[1] The simulation of the remaining benchmarks still takes 6–8 hours on a contemporary notebook.

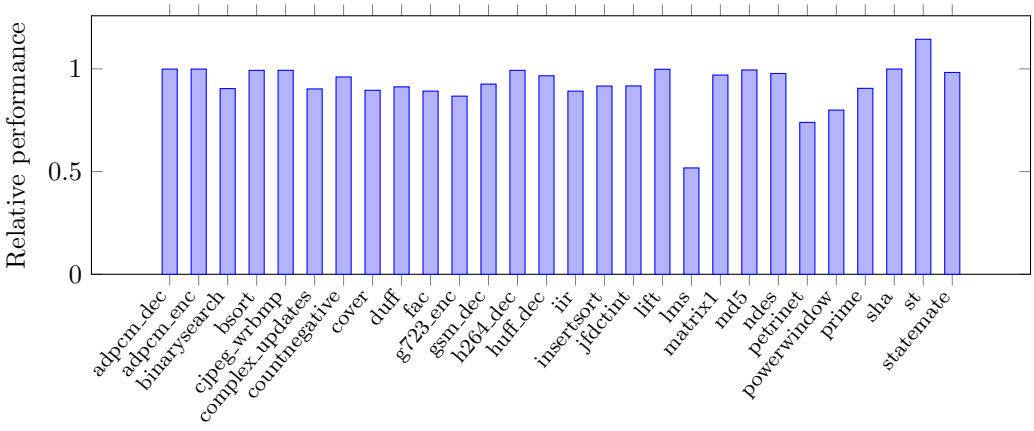

Figure 2 Relative average-case performance comparing the method cache with a standard cache on normal programs

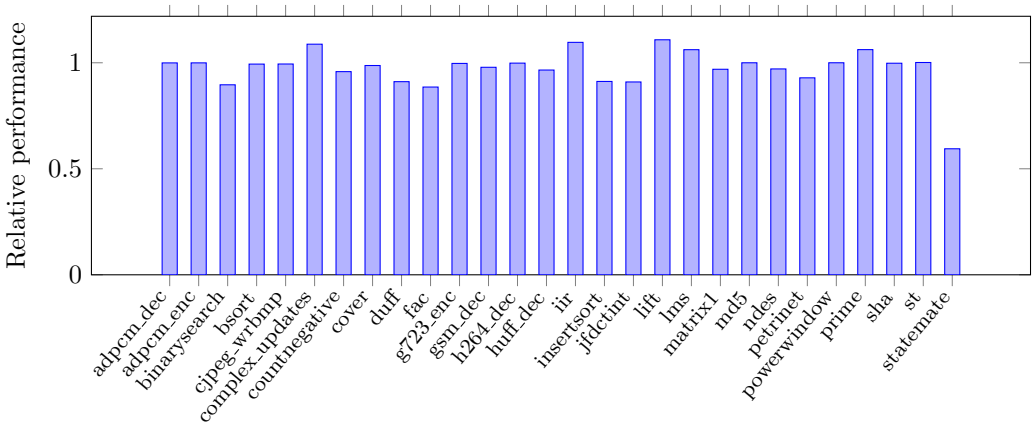

Figure 3 Relative single-path performance comparing the method cache with a standard cache

5.2 Single-Path Comparison and Prefetching

Figure 3 shows the performance comparison between a method cache and a standard cache with single-path generated code. The figure is now more diverse than the average-case figure. Some benchmarks gain and some loose when using a method cache. There is no clear winner.

Figure 4 show the performance comparison between a method cache and an instruction cache that includes the prefetching unit. The results are like the comparison in Figure 3. Some benchmarks gain a little bit with the prefetching unit. We assume that most benchmarks are almost fitting into the cache and leaving not enough room for improvement by prefetching. It has been shown that smaller caches benefit most from the prefetcher [3].

5.3 Associativity

Figure 5 shows the comparison of a 2-way cache with LRU replacement with a direct mapped instruction cache. Originally we assumed that a direct mapped cache is a better fit for single-path code as is avoids cache trashing on loops that are larger than the cache. However, we see in the figure that some benchmarks benefit from a higher associativity. Only `statemate`

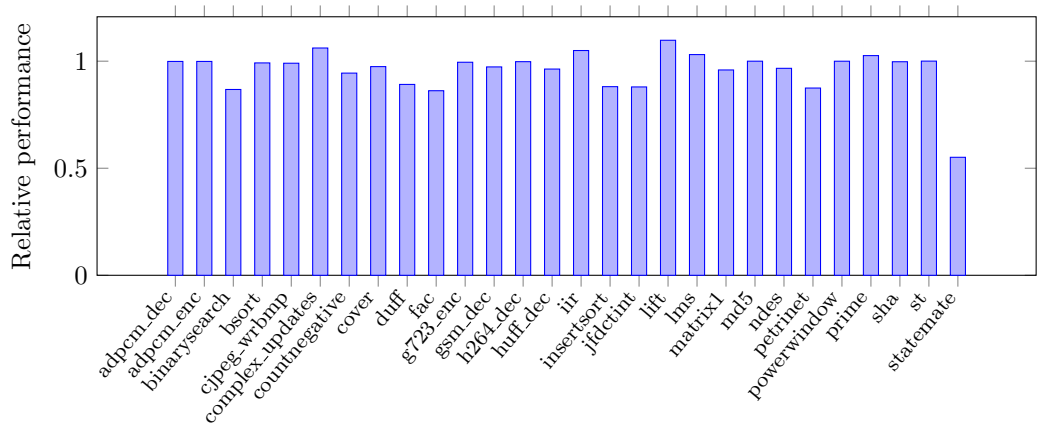

Figure 4 Relative single-path performance comparing the method cache with a prefetching cache

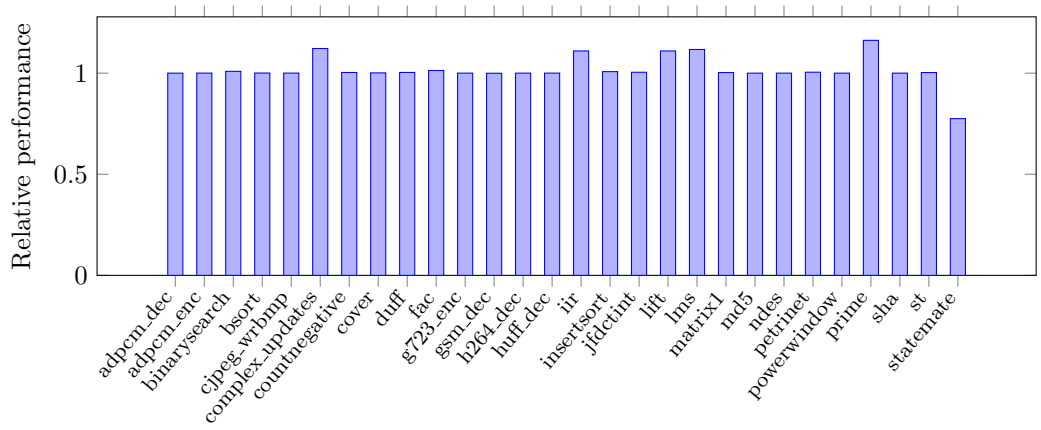

Figure 5 Relative single-path performance comparing a 2-way cache with a direct mapped cache

performs better with a direct mapped cache. Therefore, we deduct that the 4 KB of one way is large enough for the larger loops in the benchmarks.

5.4 Avoiding Function Splitting

The compiler for Patmos contains a so called "function splitter". The function splitter is in charge to split functions that are too large to fit into the method cache into sub-functions. However, the heuristics of the function splitter also tries to minimize code blocks loaded into the cache that might not get executed by splitting functions into sub-functions. However, this generates more functions and the method cache has a limit of holding at most 16 functions at the same time in the cache. Therefore, it is interesting how the method cache preforms with the original function layout of the benchmarks. For this experiment, we assume that all functions of the benchmarks are less than 8 KB.

Figure 6 show that most benchmarks benefit from *not* using the function splitter. In this comparison the method cache outperforms a standard cache in almost all cases. This is an indication that more work in the function splitter is needed to produce the best code for single-path code. Probably a feedback loop with profiling (measuring execution time) would be beneficial.

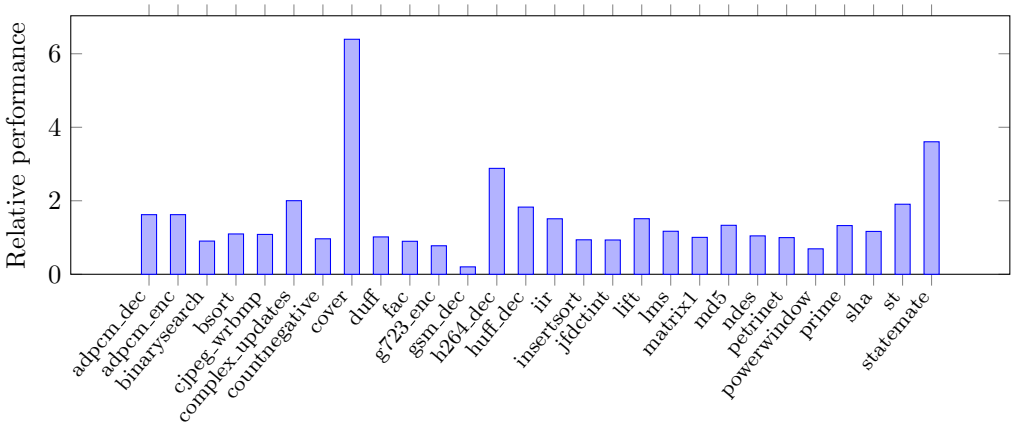

Figure 6 Relative single-path performance comparing the method cache with a standard cache without function splitting

5.5 Discussion

Single-path code has different characteristics than normal code. We see some of the different characteristics when comparing different caching methods. The method cache, which works not so well in the average-case performance, is a better fit when using single-path code. Prefetching with a standard cache provides some benefit, but not too much on an 8 KB large cache. The most surprising result is that avoiding function splitting with the method cache shows considerable improvement of using a method cache compared to a standard instruction cache. This result is promising and an indication that the function splitter needs a single-path specific heuristics. We consider adapting the function splitter for single-path code as future work.

As we see in the results, there is no clear winner for all benchmarks. Therefore, if we use an FPGA as execution platform, we can select an application specific caching method. This is like an application specific instruction set in a processor.

5.6 Reproducing the Results

We think reproducibility is of primary importance in science. As we are working in the context of an open-source project, it is relative easy to provide pointers and a description how to reproduce the presented results.

The T-CREST project is open-source and the README[2] of the Patmos repository provides a brief introduction how to setup an Ubuntu installation for T-CREST and how to build T-CREST from the source. More detailed installation instructions, including setup on Mac OS X, are available in the Patmos handbook [19]. To simplify the evaluation, we also provide a VM[3] where all needed packages and tools are already preinstalled. However, that VM is currently used in teaching and does not contain the latest version of T-CREST, including the scripts for the experiments. Therefore, you need to reinstall and build T-CREST as described in the README.

[2] `https://github.com/t-crest/patmos`
[3] `http://patmos.compute.dtu.dk/`

We have scripted all experiments and host those scripts in the `misc` repository of T-CREST. Details to rerun the experiments are described in a README.[4] The `Makefile` is setup to run the base experiments and for producing the figures as PDFs. Variations can be obtained by changing some variables.

6 Conclusion

In this paper, we compared different caching methods for single-path code. We found that the method cache, which performs to so well in the average case, shows an improvement on some benchmarks when compared to a standard instruction cache. Especially when we avoid function splitting, the method cache is the best solution for most benchmarks. This is an indication that we need a better heuristic for the function splitter for single-path code. When we use an FPGA as execution platform we have the freedom to choose the best caching solution for each individual application.

References

1 J. Allen, K. Kennedy, C. Porterfield, and J. Warren. Conversion of Control Dependence to Data Dependence. In *Proc. 10th ACM Symposium on Principles of Programming Languages*, pages 177–189, Jan. 1983.

2 Bekim Cilku, Daniel Prokesch, and Peter Puschner. A time-predictable instruction-cache architecture that uses prefetching and cache locking. In *Proc. 18th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC) Workshops, 11th IEEE/IFIP International Workshop on Software Technologies for Future Embedded and Ubiquitous Systems (SEUS)*, pages 74–79. IEEE CS Press, 2015.

3 Bekim Cilku, Wolfgang Puffitsch, Daniel Prokesch, Martin Schoeberl, and Peter Puschner. Improving performance of single-path code through a time-predictable memory hierarchy. In *Proceedings of the 20th IEEE International Symposium on Real-Time Computing (ISORC 2017)*, Toronto, Canada, May 2017. IEEE.

4 Philipp Degasperi, Stefan Hepp, Wolfgang Puffitsch, and Martin Schoeberl. A method cache for Patmos. In *Proceedings of the 17th IEEE Symposium on Object/Component/Service-oriented Real-time Distributed Computing (ISORC 2014)*, pages 100–108, Reno, Nevada, USA, June 2014. IEEE.

5 Huping Ding, Yun Liang, and Tulika Mitra. Wcet-centric dynamic instruction cache locking. In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pages 1–6. IEEE, 2014.

6 Heiko Falk, Sebastian Altmeyer, Peter Hellinckx, Björn Lisper, Wolfgang Puffitsch, Christine Rochange, Martin Schoeberl, Rasmus Bo Sørensen, Peter Wägemann, and Simon Wegener. TACLeBench: A benchmark collection to support worst-case execution time research. In Martin Schoeberl, editor, *16th International Workshop on Worst-Case Execution Time Analysis (WCET 2016)*, volume 55 of *OpenAccess Series in Informatics (OASIcs)*, pages 2:1–2:10, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

7 Heiko Falk, Sascha Plazar, and Henrik Theiling. Compile-time decided instruction cache locking using worst-case execution paths. In *Proceedings of the 5th IEEE/ACM inter-*

[4] `https://github.com/t-crest/patmos-misc/tree/master/experiments/cache_prefetch_lock/wcet2017`

national conference on Hardware/software codesign and system synthesis, pages 143–148. ACM, 2007.

8 Reinhold Heckmann and Christian Ferdinand. Worst-case execution time prediction by static program analysis. Technical report, AbsInt Angewandte Informatik GmbH. [Online, last accessed November 2013].

9 Reinhold Heckmann, Marc Langenbach, Stephan Thesing, and Reinhard Wilhelm. The influence of processor architecture on the design and the results of wcet tools. *Proceedings of the IEEE*, 91(7):1038–1054, 2003.

10 Stefan Hepp and Florian Brandner. Splitting functions into single-entry regions. In Karam S. Chatha, Rolf Ernst, Anand Raghunathan, and Ravishankar Iyer, editors, *2014 International Conference on Compilers, Architecture and Synthesis for Embedded Systems, CASES 2014, Uttar Pradesh, India, October 12-17, 2014*, pages 17:1–17:10. ACM, 2014.

11 Stefan Hepp, Benedikt Huber, Jens Knoop, Daniel Prokesch, and Peter P. Puschner. The platin tool kit - the T-CREST approach for compiler and WCET integration. In *Proceedings 18th Kolloquium Programmiersprachen und Grundlagen der Programmierung, KPS 2015, Pörtschach, Austria, October 5-7, 2015*, 2015.

12 Mingsong Lv, Nan Guan, Jan Reineke, Reinhard Wilhelm, and Wang Yi. A survey on static cache analysis for real-time systems. *Leibniz Transactions on Embedded Systems*, 3(1):05–1, 2016.

13 Daniel Prokesch, Benedikt Huber, and Peter P. Puschner. Towards automated generation of time-predictable code. In Heiko Falk, editor, *14th International Workshop on Worst-Case Execution Time Analysis, WCET 2014, July 8, 2014, Ulm, Germany*, volume 39 of *OASICS*, pages 103–112. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2014.

14 Isabelle Puaut and David Decotigny. Low-complexity algorithms for static cache locking in multitasking hard real-time systems. In *Real-Time Systems Symposium, 2002. RTSS 2002. 23rd IEEE*, pages 114–123. IEEE, 2002.

15 Peter Puschner and Alan Burns. Writing temporally predictable code. In *Proceedings of the The Seventh IEEE International Workshop on Object-Oriented Real-Time Dependable Systems (WORDS 2002)*, pages 85–94, Washington, DC, USA, 2002. IEEE Computer Society.

16 Peter Puschner, Daniel Prokesch, Benedikt Huber, Jens Knoop, Stefan Hepp, and Gernot Gebhard. The T-CREST approach of compiler and WCET-analysis integration. In *9th Workshop on Software Technologies for Future Embedded and Ubiquitious Systems (SEUS 2013)*, pages 33–40, 2013.

17 Jan Reineke, Daniel Grund, Christoph Berg, and Reinhard Wilhelm. Timing predictability of cache replacement policies. *Real-Time Systems*, 37(2):99–122, 2007.

18 Martin Schoeberl, Sahar Abbaspour, Benny Akesson, Neil Audsley, Raffaele Capasso, Jamie Garside, Kees Goossens, Sven Goossens, Scott Hansen, Reinhold Heckmann, Stefan Hepp, Benedikt Huber, Alexander Jordan, Evangelia Kasapaki, Jens Knoop, Yonghui Li, Daniel Prokesch, Wolfgang Puffitsch, Peter Puschner, André Rocha, Cláudio Silva, Jens Sparsø, and Alessandro Tocchi. T-CREST: Time-predictable multi-core architecture for embedded systems. *Journal of Systems Architecture*, 61(9):449–471, 2015.

19 Martin Schoeberl, Florian Brandner, Stefan Hepp, Wolfgang Puffitsch, and Daniel Prokesch. Patmos reference handbook. Technical report, Technical University of Denmark, 2014.

20 Martin Schoeberl, Pascal Schleuniger, Wolfgang Puffitsch, Florian Brandner, Christian W. Probst, Sven Karlsson, and Tommy Thorn. Towards a time-predictable dual-issue microprocessor: The Patmos approach. In *First Workshop on Bringing Theory to Practice: Predictability and Performance in Embedded Systems (PPES 2011)*, pages 11–20, Grenoble, France, March 2011.